The PCIe-based readout system for the LHCb experiment

J. P. Cachemiche, P.-Y. Duval, R Le Gac, F Hachon, F Réthoré

The PCIe-based readout system for the LHCb experiment

J.P. Cachemiche[a], P.Y. Duval[a], F. Hachon[a], R. Le Gac[a] and F. Réthoré[a]

[a] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*
E-mail: rethore@cppm.in2p3.fr

ABSTRACT:

The LHCb experiment is designed to study differences between particles and anti-particles as well as very rare decays in the beauty and charm sector at the LHC. The detector will be upgraded in 2019 in order to significantly increase its efficiency, by removing the first-level hardware trigger. The upgrade experiment will implement a trigger-less readout system in which all the data from every LHC bunch-crossing are transported to the computing farm over 12000 optical links without hardware filtering. The event building and event selection are carried out entirely in the farm. Another original feature of the system is that data transmitted through these fibres arrive directly to computers through a specially designed PCIe card called PCIe40.

The same board handles the data acquisition flow and the distribution of fast and slow controls to the detector front-end electronics. It embeds one of the most powerful FPGAs currently available on the market with 1.2 million logic cells. The board has a bandwidth of 480 Gbits/s in both input and output over optical links and 100 Gbits/s over the PCI Express bus to the CPU.

We will present how data circulate through the board and in the PC server for achieving the event building. We will focus on specific issues regarding the design of such a board with a very large FPGA, in particular in terms of power supply dimensioning and thermal simulations.

The features of the board will be detailed and we will finally present the first performance measurements

KEYWORDS: Detector control systems; Electronic detector readout concepts; Modular electronics.

Contents

1. LHCb New readout architecture upgrade

The LHCb collaboration will upgrade its detector in 2019 in order to run with an instantaneous luminosity multiplied by a factor 5 and with a trigger-less readout system [1]. To achieve this goal, substantial changes are required in the readout architecture.

The trigger, currently implemented in hardware and software, will migrate to a full software version running in the computer farm, once the events are assembled [2]. For each bunch crossing, data from all sub-detectors will be received by a single computing node of the farm in a round robin manner. Since all the information is available, complete event reconstruction can be run improving the selection efficiency.

This upgraded readout system is composed of five main blocks as shown in Figure 1:

- Front-ends electronics
- Event-builder Farm.
- Timing and Fast Control (TFC) supervisor [3]
- Online network
- Event-filter Farm.

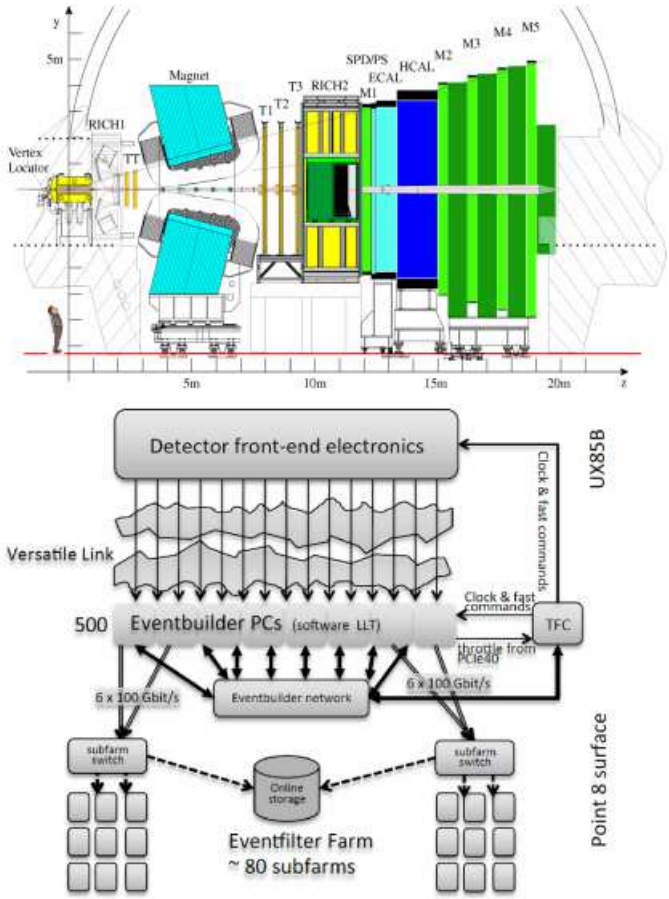


Figure 1: General architecture of the upgraded LHCb Readout system.

1.1 Rationale of the architecture

Such an architecture is called “trigger-less readout” since there is no hardware trigger filtering the data before the events building stage. Instead all event fragments are routed to the Event Builder Farm and then transmitted to Event Filter Farm where the events reconstruction and selection is performed. Figure 2 illustrates the concept of “trigger-less readout”:

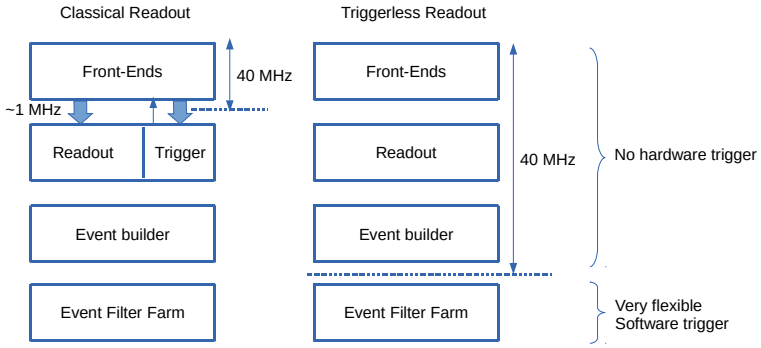


Figure 2: Classical readout vs trigger-less readout

More than ten years separate the existing LHCb readout system from the one that will be implemented for the upgrade. During that time, the power of CPUs has been multiplied by more than a factor 30. A full event was built at 1 MHz by the CPUs and online network of 2008. Ten years later, the same operation can process all the LHC collisions at an affordable cost.

1.2 New readout scheme

The LHCb collaboration takes advantage of most recent technological breakthroughs in computing to bring interesting solutions in terms of memory resources and cost optimization. They are based on PCI Express bus GEN3, its interconnection to CPU and on bi-directional switch running at 100 Gbits/s.

Starting from the Ivy Bridge architecture from Intel and with generations onwards, CPUs are no longer a bottleneck for data circulation inside a PC-server. It is possible to address directly the PC-server internal memory with DMAs (Direct Memory Access) driven by PCI Express busses. Measurements [4] showed that a sustained bandwidth of 100 Gbits/s was possible for transferring data to and from the internal memory and between memory and network card.

By placing back-end boards directly inside the farm computer, the available bandwidth permits both the readout and the event-building while using the large memory of the PC as storage buffer.

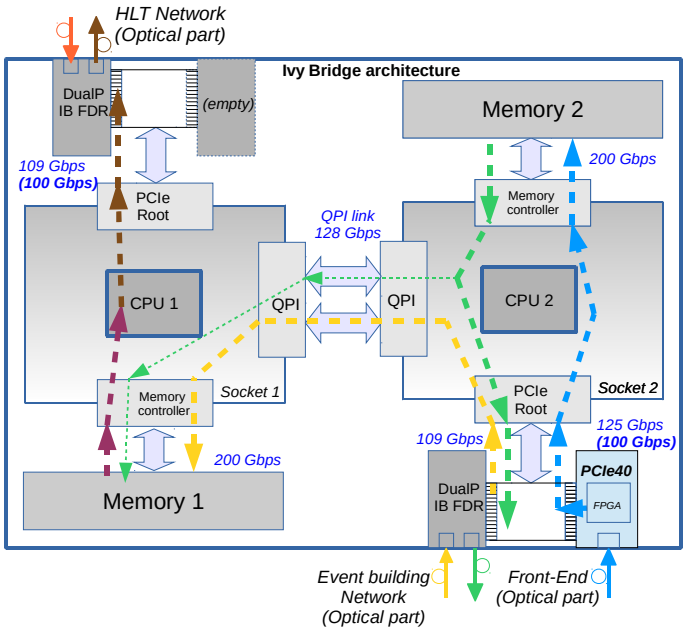
The advantages of the architecture are the following:

- intermediary crates are no longer necessary to house the readout board;
- Short distance connection between the readout boards and the PC-servers;
- Simplifies the readout board that does not embed any memory since the memory buffering is achieved by PC-servers.

1.3 Data path into computer

The feasibility of the architecture relies on sufficient bandwidth all along the data path in the PC-servers. It is estimated as 100 Gbits/s. This rate is obtained from the current dimensions of the system. With up to 12000 links [5] and an average of 24 links per card, about 500 boards are needed to perform the event building. Each link transfers from 80 to 112 bits per collision. The number of collisions per second is 30 million since the TFC will remove collision corresponding to empty crossing in the LHC frame. Therefore, we have a maximum of 40 Tbits/s that are distributed over 500 boards, corresponding to an individual data flow of 81 Gbits/s per PC-servers. This can be rounded to 100 Gbits/s with the overheads. The above figures may slightly vary but the order of magnitude stays the same.

A PC-server has generally two CPUs with their own memory banks and PCI Express interfaces. The two CPUs are tightly connected to each other through a high bandwidth QPI *(Quick Path Interconnect)* bus.



.

Figure 3: Data path into computer

Figure 3 illustrates data path into the computer. In this architecture one CPU is devoted to event building whereas the second one is devoted to the event reconstruction and selection.
Data corresponding to an event fragment are received from the FE electronics by the PCIe40 board. The PCIe40 FPGA sends the data into the Memory2 of the computer through the PCI Express bus via a DMA transfer. The event fragment is then transmitted to the target PC-server through a high speed network, like Infiniband [6], working at the same speed as incoming data *i.e.* 100 Gbits/s. The role of CPU2 is mainly to arm DMA transfers.
Each PC-server is also a target for a given bunch crossing. It receives all event fragments of a single event through the bi-directional high speed network card. Data are routed by DMA transfer to Memory1 through the QPI bus.
The first step of the event selection might be run by the CPU1. Accepted events are then sent to the Event Filter Farm where the full reconstruction and selection will be performed.

2. Full scale prototype

2.1 Main features

A full scale prototype of the readout card has been designed.
The PCIe40 card is a PCI Express board designed to interface the FE electronics to the event builder. It embeds one of the most powerful FPGAs currently available on the market (an Arria10 10AX115S4F45 from Altera) with 1.2 million logic cells and 72 High Speed Serial Links running at 10 Gbits/s or more.
When used for data acquisition, the role of the card is to merge data belonging to a single collision. To decrease the number of optical links, data are compressed in the front-ends. It is therefore not possible to compress them anymore on the PCIe40 or only in a marginal way. For this reason the output bandwidth must be roughly equal to the input bandwidth. Therefore the

100 Gbits/s bandwidth of PCI Express limits the number of input links. It is equal to 24 when the GBT [7] protocol mode is used in wide bus. However, in parts of the detector where the occupancy is low, it would be possible to concentrate more links optimizing the cost. For that reason, the PCIe40 board is equipped with 48 bidirectional optical links running at up 10 Gbits/s. In addition, the board embeds a SFP+ optical link for the TFC interface.
The card can also be used for Timing, Fast control distribution and Slow Control interface by simply reprogramming the FPGA. In this case, the card can handle up to 48 FE boards.
Figure 4 illustrates the synoptic functionality of the PCIE40 board.

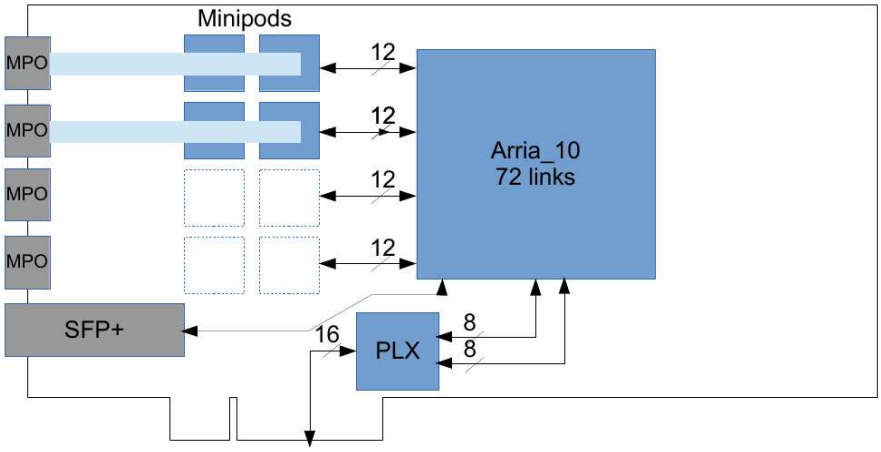


Figure 4: PCIe40 board – Synoptic drawing.

2.2 Flexibility

The overall architecture of the readout system is shown on Figure 5. The PCIe40 plays a generic role in the system. By placing the appropriate firmware, the PCIe40 can be turned into:

- A TFC supervisor, receiving the LHC clock and distributing Fast Commands and Clocks to PCIe40 cards controlling front-end boards. The link with these cards can be either the 48 minipod links in which case 12 TFC cards are necessary to drive the system, or a 1:128 PON device installed in the SFP+ cage in which case only 5 TFC cards are necessary.
- TFC and Slow Control interface cards. They receive TFC commands through the SFP+ or PON device and broadcast them over the 48 minipods fibers to front-end electronics.
- DAQ boards, processing in average 24 optical links each.

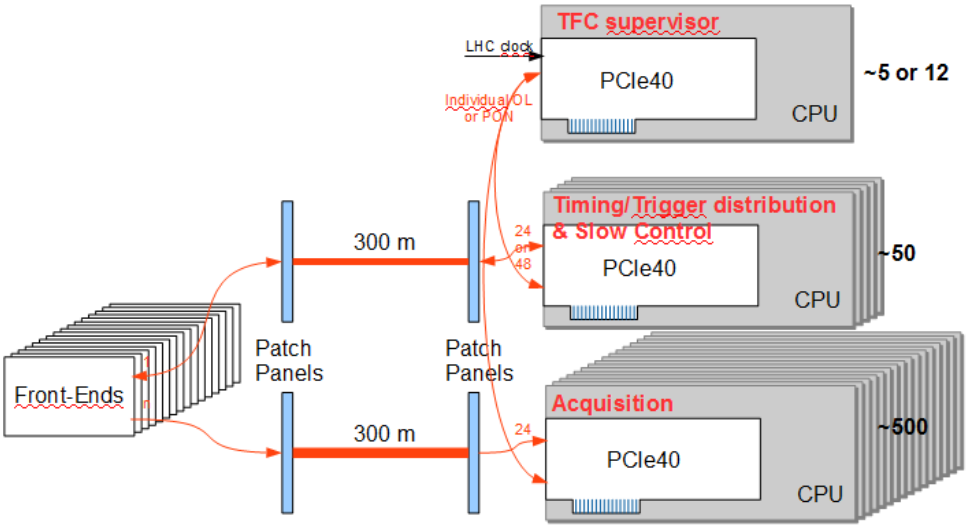


Figure 5: Overall architecture

2.3 Design of the board

Designing an electronic board compliant to PCI Express GEN3 [8] and using a high-density and high performance FPGA like the Arria10 component is really a challenging task. Its implementation requires following stringent manufacturer's recommendations to successfully meet design requirements.

The most critical points we had to take care were mainly:

- The constraints imposed by the PCI express standard in GEN3x16 configuration (lane ordering, placement on FPGA pins, very accurate line routing).
- The design of very sensitive 130 high speed serial links running up to 10 Gbits/s.
- And especially the design of the power supply providing different voltages with very high densities of current (for example up to 60A for FPGA core).

FPGA power consumption must be estimated accurately. It conditions the design of power supplies, voltage regulators, decoupling, heat sink and cooling system. ALTERA provides a spreadsheet tool named "PowerPlay EPE" that allowed us to make an early estimate of the power consumption.

2.3.1 Power Consumption estimation:

The estimation is obtained by feeding this tool with a configuration file describing all the interfaces of the FPGA on the board. Various parameters like frequency, toggle rate, memory and logic resources utilization can be tuned to provide an estimation of power consumption of each voltage used by the FPGA.

Authorizing the FPGA to run at its limits in term of resource occupancy, toggle rate and frequency requires a cumbersome number of power supply components and exposes us to cooling issues. If we are to stay in reasonable dimensions for the card and limit power consumption, we need to find a compromise.

In the absence of the final firmware the following estimations have been made to estimate the power consumption of the card:

- **Data Acquisition (DAQ):** Toggle Rate 50 % - Frequency 250 MHz - Occupancy 93%
- **Slow Control (SC) :** Toggle Rate 12.5 % - Frequency 125 MHz – Occupancy 7%
- **High speed transceivers (XCVR)**: 49 links at 4.8 Gbits/s and 16 links PCIe at 8 Gbits/s.

Figure 6 illustrates the EPE results obtained for the above configuration:

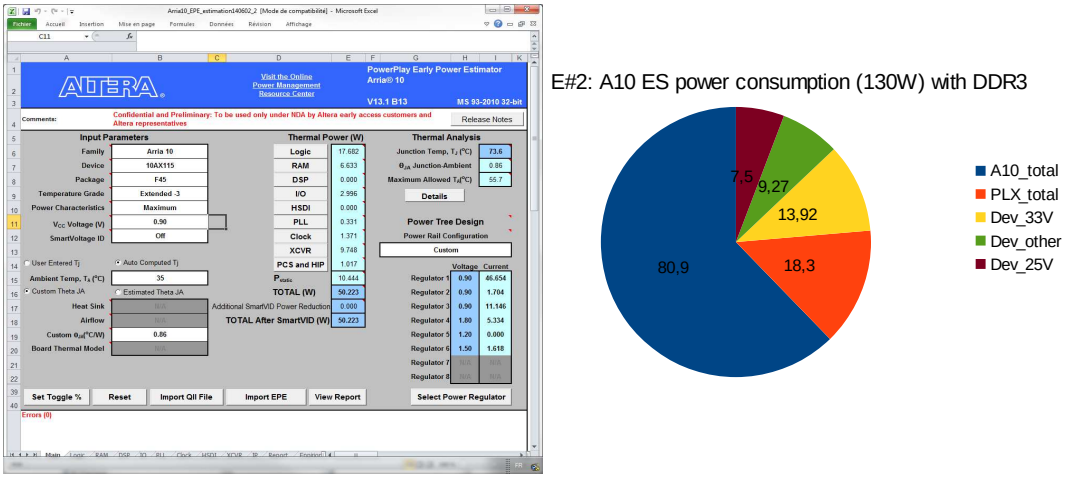


Figure 6: EPE power supply estimation

The result obtained with the EPE simulation allows the sizing of the overall power and the power distribution that are required for the FPGA. However, the EPE gives only a raw estimation of the power consumption with a precision in the order of ± 20 %. Only a simulation with the final firmware will give more accurate results. The risk of under-estimation is however limited because this estimation was made with Engineering Sample devices that are consuming around 30% more than the final production devices, according to ALTERA.

2.3.2 Power DC and temperature simulations:

Due to the high density of components on the PCIe40 board, the power supply is placed on mezzanine cards. The PCIe40 is equipped with two types of mezzanine cards. The FPGA core is supplied by the first type (0.95 V – 60 A), while the other voltages (3.3 V, 2.5 V, 1.8 V, 0.9 V) of the PCIe40 board are supplied by four mezzanine cards of the second type. In this case different voltages are obtained on similar cards by individually changing passive components on them.

The high density of the current implied on this design requires a methodical checking of each part of the PCB that links the current. Power DC simulations are compulsory to detect voltage drops on the layers but also to detect hot points that could lead to break down of the copper tracks or, in the worst case, set fire to the card.

The temperature of the FPGA in operational mode and the power dissipation of the PCB is also a crucial issue.

To avoid these problems we performed simulations checking the current density and temperature in the PCB. A temperature of 130 °C is the Maximum Operating Temperature (MOT) value allowed before meeting delamination issues for the material we used. By security we targeted a maximum value of 100 °C in our simulations.

Figure 7 gives an example of simulations performed on a power mezzanine board and of the geometrical correction we made to limit the density of currents.

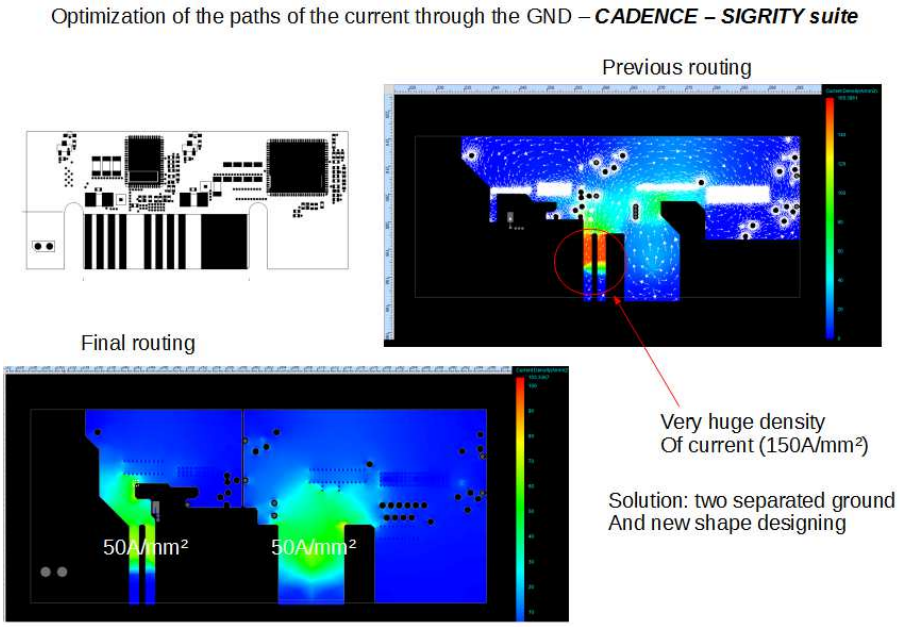


Figure 7: simulation of power DC on board

2.3.3 Prototype

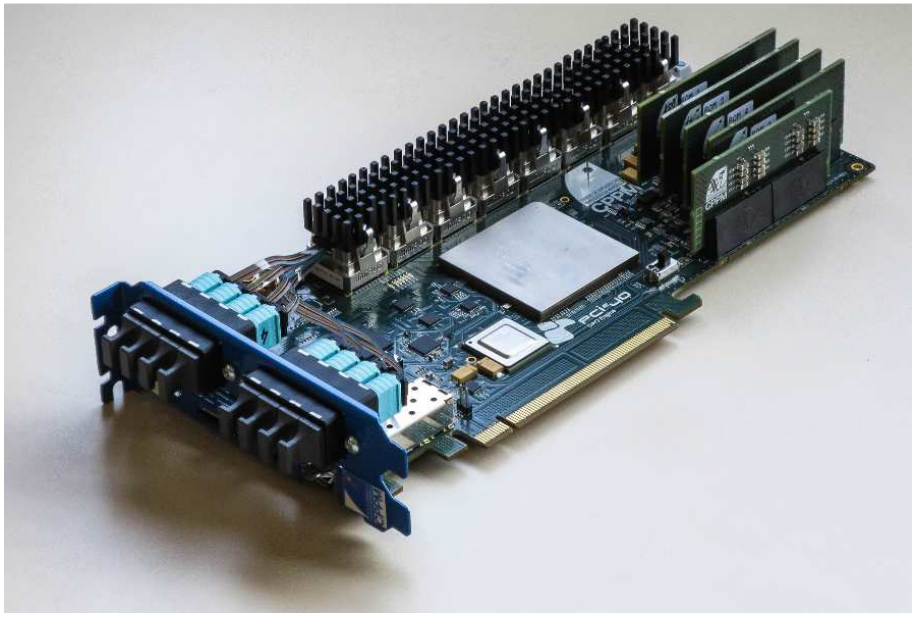


Figure 8: PCIE40 board

The PCIe40 prototype is shown in Figure 8. It is equipped with mini-pods, optical receivers, and transmitters from AVAGO. Optical drivers are pluggable which facilitates tuning the card with the optimum number of optical devices required by the aforementioned configurations.
Up to 48 optical links are available through 8 MPO connectors and one through the SFP+ interface on the front plate.

2.4 First measurements

Two PCIe40 cards equipped with Arria10 Engineer Samples (ES1&ES2) have been manufactured. The cards are fully operational, the PCIe Gen3x16 was tested on the ES2 version of the PCIe40 and all the optical links are operational and successfully tested at 10 Gbits/s.
Figure 9 illustrates a BER (Bit Error Rate) measurement for the TX part. Measurements have been made at 5 Gbits/s and 10 Gbits/s over a 10 meters OM3 optical fiber.

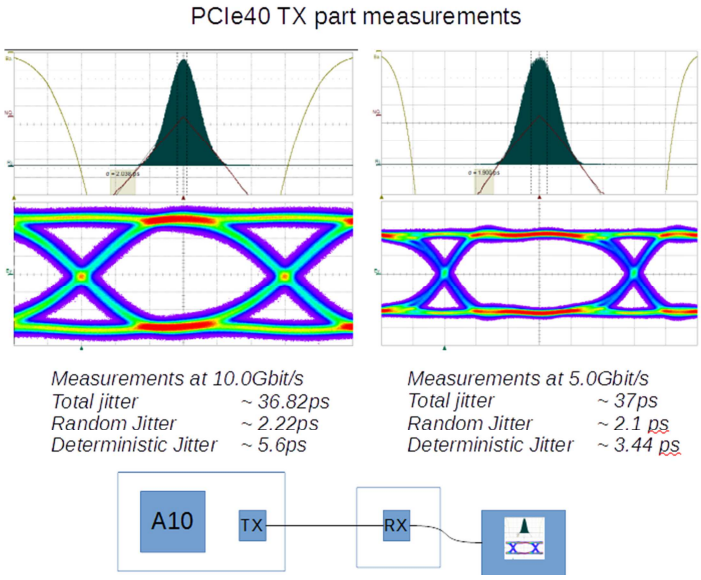


Figure 9: BER measurement

3. Conclusion

We have shown the advantages of a triggerless architecture where event building is achieved in the farms. It relies on a key element: a PCIe board able to absorb and process 100 Gbits/s of data with a very tight connection with CPUs.
Building such a board is a real challenge:

- demanding requirements in term of power supplies and cooling;
- signal integrity issues with many links running at up to 10 Gbits/s;
- complex PCIe Gen3 interface running at full speed (100 Gbits/s);
- complex PCB routing with very dense FPGA (1932 pins).

A prototype of such a board called PCIe40 has been designed and tested. The card is sufficiently flexible and reconfigurable to instantiate several functions of the system: data acquisition, TFC supervisor and TFC/Slow control interface.

The card has been successfully interfaced at full speed with a candidate PC-server. It has also been tested with the GBTx communication chip that will equip all the front-ends. The hardware feasibility of a PCIe readout card enabling the implementation of the above architecture can be considered as demonstrated.
Two prototypes of the PCIe40 board have been developed with early engineering samples of the FPGA. A small production of 20 boards is about to be launched in order to provide "miniDAQs" to test FE electronics in real conditions.
Full production is foreseen in 2016/2018 for a total of about 600 boards including spares.

Acknowledgments

We wish to thank the LHCb online team from CERN with whom we tightly collaborated in this development.

References

1. LHCb Collaboration, *Expression of Interest for an LHCb Upgrade*, CERN/LHCC/2008-007, LHCb 2008-019, 22nd April 2008. Document.
2. LHCb Collaboration, *Framework TDR for LHCb upgrade*, CERN/LHCC/2012-007, LHCb-TDR-12, 26th April 2012.
3. F. Alessio et al., *System-level specifications of the Timing and Fast Control System for the LHCb Upgrade*, CERN-LHCb-2012-001, January 2012.
4. Perez, D.H.C.; Schwemmer, R.; Neufeld, N. *Protocol-independent event building evaluator for the LHCb DAQ system* Real Time Conference (RT), 2014 19th IEEE-NPSS
5. J.-P. Cachemiche et al., *The LHCb Readout for the Run 3*, February 2014
6. InfiniBand℠ Trade Association, *Infiniband™ Architecture Specification*, Release 1.2.1, November 2007.
7. P. Moreira, *The GBTX link interface ASIC*, V1.7 Draft, 3 November 2011, Document.
8. PCI-SIG, *PCI express Base Specification*, Revision 3.0, November 10, 2010